"JUST THE MATHS"

UNIT NUMBER

18.4

STATISTICS 4
(The principle of least squares)

by

A.J.Hobson

UNIT 18.4 - STATISTICS 4

THE PRINCIPLE OF LEAST SQUARES

18.4.1 THE NORMAL EQUATIONS

Suppose two variables, x and y, are known to obey a **"straight line law"** of the form $y = a + bx$, where a and b are constants to be found.

Suppose also that, in an experiment to test this law, we obtain n pairs of values, (x_i, y_i), where i = 1,2,3,...,n.

If the values x_i are **assigned** values, they are likely to be free from error, whereas the **observed** values, y_i will be subject to experimental error.

The principle underlying the straight line of **"best fit"** is that, in its most likely position, the sum of the squares of the y-deviations, from the line, of all observed points is a minimum.

Using partial differentiation, it may be shown that

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i \quad ---(1)$$

and

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \quad ---(2).$$

The statements (1) and (2) are two simultaneous equations which may be solved for a and b.

They are called the **"normal equations"**.

A simpler notation for the normal equations is

$$\Sigma\, y = na + b\Sigma\, x$$

and

$$\Sigma xy = a\Sigma x + b\Sigma\, x^2.$$

By eliminating a and b in turn, we obtain the solutions

$$a = \frac{\Sigma\, x^2.\Sigma\, y - \Sigma x.\Sigma\, xy}{n\Sigma\, x^2 - (\Sigma\, x)^2} \quad \text{and} \quad b = \frac{n\Sigma\, xy - \Sigma\, x.\Sigma\, y}{n\Sigma\, x^2 - (\Sigma\, x)^2}.$$

With these values of a and b, the straight line $y = a + bx$ is called the **"regression line of y on x"**.

EXAMPLE

Determine the equation of the regression line of y on x for the following data which shows the Packed Cell Volume, $x$mm, and the Red Blood Cell Count, y millions, of 10 dogs:

x	45	42	56	48	42	35	58	40	39	50
y	6.53	6.30	9.52	7.50	6.99	5.90	9.49	6.20	6.55	8.72

Solution

x	y	xy	x^2
45	6.53	293.85	2025
42	6.30	264.60	1764
56	9.52	533.12	3136
48	7.50	360.00	2304
42	6.99	293.58	1764
35	5.90	206.50	1225
58	9.49	550.42	3364
40	6.20	248.00	1600
39	6.55	255.45	1521
50	8.72	436.00	2500
455	73.70	3441.52	21203

The regression line of y on x thus has equation $y = a + bx$, where

$$a = \frac{(21203)(73.70) - (455)(3441.52)}{(10)(21203) - (455)^2} \simeq -0.645$$

and

$$b = \frac{(10)(3441.52) - (455)(73.70)}{(10)21203) - (455)^2} \simeq 0.176$$

Thus, $y = 0.176x - 0.645$

18.4.2. SIMPLIFIED CALCULATION OF REGRESSION LINES

A simpler method of determining the regression line of y on x for a given set of data, is to consider a temporary change of origin to the point $(\overline{x}, \overline{y})$, where $\overline{x}$ is the arithmetic mean of the values x_i and $\overline{y}$ is the arithmetic mean of the values y_i.

RESULT

The regression line of y on x contains the point $(\overline{x}, \overline{y})$.

Proof:

From the first of the normal equations,

$$\frac{\Sigma\ y}{n} = a + b\frac{\Sigma\ x}{n}.$$

That is,

$$\overline{y} = a + b\overline{x}.$$

A change of origin to the point $(\overline{x}, \overline{y})$, with new variables X and Y, is associated with the formulae

$$X = x - \overline{x} \ \text{ and } \ Y = y - \overline{y}$$

and, in this system of reference, the regression line will pass through the origin.

Its equation is therefore

$$Y = BX,$$

where

$$B = \frac{n\Sigma\ XY - \Sigma\ X.\Sigma\ Y}{n\Sigma\ X^2 - (\Sigma\ X)^2}.$$

However,

$$\Sigma\ X = \Sigma\ (x - \overline{x}) = \Sigma\ x - \Sigma\ \overline{x} = n\overline{x} - n\overline{x} = 0$$

and

$$\Sigma\ Y = \Sigma\ (y - \overline{y}) = \Sigma\ y - \Sigma\ \overline{y} = n\overline{y} - n\overline{y} = 0.$$

Thus,

$$B = \frac{\Sigma\ XY}{\Sigma\ X^2}.$$

Note:
In a given problem, we make a table of values of x_i, y_i, X_i, Y_i, X_iY_i and X_i^2.

The regression line is then

$$y - \overline{y} = B(x - \overline{x}) \quad \text{or} \quad y = BX + (\overline{y} - B\overline{x});$$

though, there may be slight differences in the result obtained compared with that from the earlier method.

EXAMPLE

Determine the equation of the regression line of y on x for the following data which shows the Packed Cell Volume, $x$mm, and the Red Blood Cell Count, y millions, of 10 dogs:

x	45	42	56	48	42	35	58	40	39	50
y	6.53	6.30	9.52	7.50	6.99	5.90	9.49	6.20	6.55	8.72

Solution

The arithmetic mean of the x values is $\overline{x} = 45.5$

The arithmetic mean of the y values is $\overline{y} = 7.37$

This gives the following table:

x	y	$X = x - \overline{x}$	$Y = y - \overline{y}$	XY	X^2
45	6.53	-0.5	-0.84	0.42	0.25
42	6.30	-3.5	-1.07	3.745	12.25
56	9.52	10.5	2.15	22.575	110.25
48	7.50	2.5	0.13	0.325	6.25
42	6.99	-3.5	-0.38	1.33	12.25
35	5.90	-10.5	-1.47	15.435	110.25
58	9.49	12.5	2.12	26.5	156.25
40	6.20	-5.5	-1.17	6.435	30.25
39	6.55	-6.5	-0.82	5.33	42.25
50	8.72	4.5	1.35	6.075	20.25
455	73.70			88.17	500.5

Hence,

$$B = \frac{88.17}{500.5} \simeq 0.176$$

and so the regression line has equation

$$y = 0.176x + (7.37 - 0.176 \times 45.5).$$

That is,

$$y = 0.176x - 0.638$$

18.4.3 EXERCISES

1. For the following tables, determine the regression line of y on x, assuming that $y = a+bx$.

(a)

x	0	2	3	5	6
y	6	-1	-3	-10	-16

(b)

x	0	20	40	60	80
y	54	65	75	85	96

(c)

x	1	3	5	10	12
y	58	55	40	37	22

2. To determine the relation between the normal stress and the shear resistance of soil, a shear-box experiment was performed, giving the following results:

Normal Stress, x p.s.i.	11	13	15	17	19	21
Shear Stress, y p.s.i.	15.2	17.7	19.3	21.5	23.9	25.4

If $y = a + bx$, determine the regression line of y on x.

3. Fuel consumption, y miles per gallon, at speeds of x miles per hour, is given by the following table:

x	20	30	40	50	60	70	80	90
y	18.3	18.8	19.1	19.3	19.5	19.7	19.8	20.0

Assuming that

$$y = a + \frac{b}{x},$$

determine the most probable values of a and b.

18.4.4 ANSWERS TO EXERCISES

1. (a)

$$y = 6.46 - 3.52x;$$

(b)

$$y = 54.20 + 0.52x;$$

(c)

$$y = 60.78 - 2.97x.$$

2.

$$y = 4.09 + 1.03x.$$

3.

$$a \simeq -42 \quad \text{and} \quad b \simeq 20.$$